Fibre Channel over Convergence Enhanced Ethernet Enterprise Data Center Use Cases

Daniel Eisenhauer and Renato Recio

***Abstract*— The traditional data center (DC) compute model, especially in the x86 space, has consisted of lightly utilized servers running a bare metal OS or a Hypervisor with a small number of Virtual Machines. In this traditional model, servers attach to the network lower bandwidth links, such as 1 Gbps Ethernet and 2 or 4 Gbps Fibre Channel. The physical compute model suffers from two major issues: High capital expenses due to under utilized servers and multiple fabrics; and High operational expenses due to manual administration of many management tools.**

We see the industry moving to a Dynamic Infrastructure Networking model that has highly utilized servers running many VMs per server and uses high bandwidth links to communicate with virtual storage and virtual networks. This paper will describe seven evolutionary use cases towards this new model. It will describe how they lead to: Lower capital expenses through: higher utilization (server, storage and network), and converged fabrics; and Lower operational expenses through automated and integrated management that optimizes data center infrastructure.

***Index Terms*—Convergence Enhanced Ethernet (CEE), Fibre Channel over Convergence Enhanced Ethernet (FCoCEE), InfiniBand (IB), iSCSI, SAN, NAS**

I. Introduction

Today customers have three options for fabric convergence: InfiniBand, NAS (Network Attached Storage) over Ethernet and iSCSI. Though InfiniBand has been accepted in High Performance Computing, it is not a DCN convergence contender. It cannot provide full fabric convergence, because Ethernet is still needed, and doesn't have tier-1 storage vendor support. iSCSI was introduced in 2001 and after a few false starts, is beginning to climb up the volume S-curve, but it lags in performance and lacks Enterprise class functions. NAS's simplification qualities (e.g. file based management) have fueled its SMB and middle-tier server adoption. Scale-out NAS and Enterprise class function enhancements (e.g. remote site back-up) are making it a formidable fabric convergence alternative.

Enterprise customers that require Enterprise class storage and have a significant FC install base (i.e. many of our customers) would like to leverage their FC investment (hardware, management tools, and skills), while at the same time reaping the value proposition of fabric convergence. Two recent industry initiatives seek to satisfy this need: Convergence Enhanced Ethernet (CEE) and FC over Ethernet (a.k.a. FCoE, FCoCEE).

1) Convergence Enhanced Ethernet[1] (CEE) improves Ethernet's ability to carry multiple traffic flows and provide multiple paths between endpoints. The CEE Authors have published version 0 specification proposals to IEEE for these enhancements, which also enable vendors to build de-facto standard CEE component implementations.

2) FC over Ethernet[1] (a.k.a. FCoE, FCoCEE) replaces FC's physical and transmission layers with Ethernet. By keeping FC's framing and upper layers the same, FCoE preserves the investments made in FC infrastructure (e.g. management and operating system stacks). The T11 standard organization is expected to publish the FCoE specification (formally known as FC-BB-5) in 2H/2009.

This paper will describe use cases that allow Enterprise clients to take full advantage of convergence savings, in an evolutionary manner that mitigates the risks associated with fabric convergence. It will also describe areas where further work is needed to fully realize the value of converged and virtualized Ethernet switching.

II. Towards a Converged Data Center

As shown in figure 1 below, the value of fabric convergence is clear. Using multiple dedicated fabrics requires separate components, adapters, cables, switches, and fabric management for each fabric type. Whereas carrying multiple traffic types over a single converged fabric potentially reduces hardware, energy and management costs. This component elimination also improves reliability, because it reduces the number of: failure points and cables that can be

D. G. Eisenhauer is with IBM, 11400 Burnet Rd., Austin, TX 78758 USA (e-mail: dge@us.ibm.com)

R. J. Recio is with IBM, 11400 Burnet Rd., Austin, TX 78758 USA (e-mail: recio@us.ibm.com)

wrongly configured. It also offers the potential of simplifying the management through the use of a single management console.

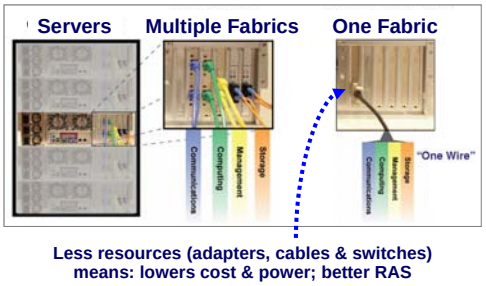

Figure 1

Fabric convergence also faces some adoption barriers that must be overcome. Most Enterprise data centers have different organizations for managing servers, storage (including storage networks) and IP/Ethernet networks. Fabric convergence requires the storage and IP/Ethernet organizations to not just work closely together, but to co-ordinate fabric configuration, provisioning, orchestration and monitoring. This obviously cannot be done without support from the underlying management tools. That is, the underlying tools need to support discovery, monitoring and configuration of the Quality of Service CEE capabilities, as well as the FC and FCoE capabilities. Initial adoption using the de-facto standard versions of the Ethernet enhancements (i.e. CEE) and the near final standard version of FCoE are just now coming to market[2]. Standard, technology and product maturity is another factor that must be taken into account when considering Enterprise production deployments of this new technology.

The following sections describe an evolutionary approach towards data center wide fabric convergence. The rate of progressing through these evolutionary steps is correlated with the rate at which the above adoption barriers are addressed.

A. Chassis level fabric convergence

Today's rack optimized servers use Fibre Channel (FC) adapters to attach an FC Top-of-Rack (TOR) switch, which connects to the data center's FC fabric. For Ethernet, today's rack optimized servers use Ethernet adapters to attach an Ethernet Top-of-Rack (TOR) switch, which connects to the data center's Ethernet fabric.

Similarly, in today's blade servers, Fibre Channel (FC) adapters are connected to the data center's FC fabric through an integrated blade FC switch. For Ethernet, the server's Ethernet adapters are used to attach an integrated blade Ethernet switch, which connects to the data center's Ethernet fabric.

As described in the previous section, if additional fabrics are used for management and cluster communications, those fabrics require additional adapters and switches.

Figure 2 below depicts the configurations described above.

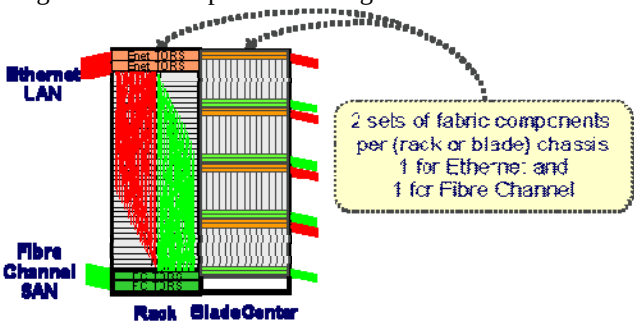

Figure 2

Under use case (A) fabric convergence is contained to a single chassis, which may be a Blade Server chassis or a rack level chassis. Within the (rack or blade) chassis a converged network adapter[2] (CNA) is used to connect to a Fibre Channel over Ethernet Forwarder enabled switch, which uses Ethernet and Fibre Channel to connect to the data centers existing Ethernet and Fibre Channel fabrics, respectively. By eliminating the need for Fibre Channel adapters and switches within the chassis, this approach reduces the number of adapters and switches within the chassis by half.

Figure 3 below depicts converged use case (A) for both a Rack and Blade level chassis. As an example of the cost savings, using an FCoE enabled Top-of-Rack (TOR) switch[3] for Rack optimized servers with a single rack unit form factor, the configuration below eliminates over 36 FC adapters and two FC top-of-rack switches.

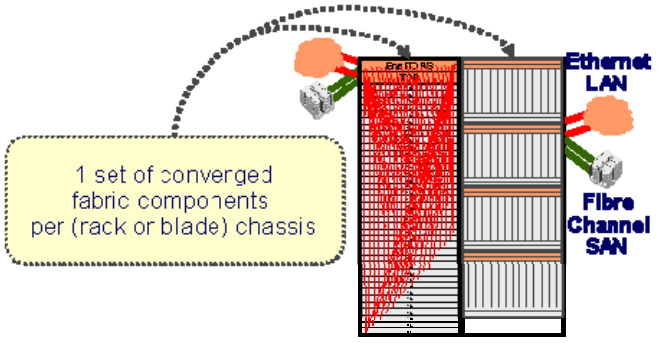

Figure 3

In order to deploy configuration shown in figure 3 in a production environment, the amount of bandwidth that needs to be allocated for each traffic classes must be known. In many cases, these bandwidth allocations are not static,

especially for virtualized environments that use Virtual Machine migration to increase server utilization. Even if the environment is not virtualized, bandwidth allocations are still likely to be dynamic. For example, the amount of bandwidth allocated to a storage traffic class is likely to be higher during server data save windows than under normal operation. In net, the network change and configuration management tools must provide the ability to dynamically change the bandwidth allocated to each traffic class. Similar to today's integrated chassis switches (e.g. a Blade switch), this use case contains converged switch management to the chassis level. This approach contains the management barriers to the chassis level versus having to solve the problem at the data center level.

Application workload modeling can be used to project the bandwidth allocation per traffic class. However, to further ease overcoming the management barriers described earlier, we recommend the configuration in figure 3 first be deployed in a development/test environment, in order to determine exactly how much bandwidth to allocate per traffic class over the course of business operations. Once the bandwidth per traffic class is understood, use case (A) can be deployed in production environments. Obviously, monitoring the configuration's performance is still required, so that additional dynamic bandwidth allocation adjustments can be made.

We expect the server's existing virtualization infrastructure (e.g. a Hypervisor) to be extended to support use case (A). This creates two sub-cases for use case (A):

1) Indirectly shared FCoCEE CNAs

Under this sub-case an FCoCEE capable Converged Network Adapters (CNAs) is shared through extensions to the server's existing virtualization infrastructure[4].

Today's server virtualization infrastructure uses an integrated Virtual Ethernet Bridge (VEB) for communication with external systems and between local Virtual Machines. For this use case, each VM uses this same VEB to share the underlying FCoCEE CNA.

For the FC path, today's FC devices are attached to the server virtualization infrastructure, which exports one or more of these devices to local VMs.

2) Directly shared FCoCEE CNAs

Under this sub-case an FCoCEE capable Converged Network Adapters (CNAs) can be directly shared by multiple local VMs[5]. With the advent of PCIe adapters supporting multi-queue, multi-function or Single-Root IO Virtualization (SR-IOV), enterprise class methods for directly sharing IO are becoming available for x86, high volume servers. These virtualization approaches enable a Virtual Machine's device driver to bypass the Hypervisor and thereby directly share a single PCIe adapter across multiple Virtual Machines (VMs).

For the Ethernet path, these CNAs support an integrated Virtual Ethernet Bridge[6] (VEB) for communication with external systems and between local Virtual Machines. The CNA may also support the ability to use an external switch for local VM-VM bridging.

For the FC overlay path, the CNA uses FC device IO virtualization to provide each VM its own virtual FC host bus adapter.

B. Large SMP level fabric convergence

Large SMPs, such as a Power 750 server, use a server virtualization infrastructure (e.g. a Hypervisor) to consolidate many Virtual Machines (VMs) onto a single server. The server virtualization infrastructure uses an integrated Virtual Ethernet Bridge (VEB) for communication with external systems and between local Virtual Machines. The integration of many VMs into a large SMP, essentially converts the VEB into an access layer switch, which eliminates the need for a standalone aggregation switch, such as a TOR switch.

Figure 4 below depicts the value of fabric convergence for a large SMP that uses multiple adapters to connect to the data center's existing Ethernet and FC fabrics.

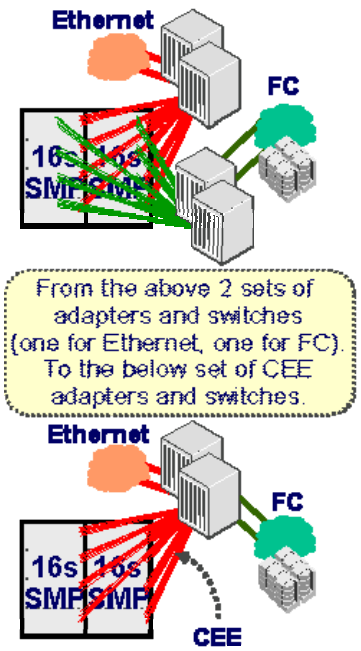

Figure 4

As shown in the top of figure 4, a set of Ethernet adapters is used to connect the large SMP a Modular Ethernet switch, which is connected to the data center's Ethernet infrastructure. Similarly, a set of FC adapters is used to connect the large SMP a Modular FC switch, which is connected to the data center's FC infrastructure.

On the bottom of figure 4 is a large SMP that uses FC over CEE enabled CNAs to attach to a converged Modular switch. This switch is used to connect into the data center's existing Ethernet and FC infrastructures. For the same reasons as described in use case (A), we recommend the configuration in figure 4 be first deployed in a development/test environment.

C. Chassis level Cloud Building Block fabric convergence

A Cloud Building Block (CBB) describes a data center granular unit of scale, which includes the integrated servers, storage and network equipment, as well as the virtualization infrastructure and the associated platform and service management functions.

This use case extends the chassis level convergence model described in (A) and (B) to a CBB of interconnected servers. That is, a converged fabric is used within each rack or blade chassis, but at the CBB level there are still two separate fabrics: Ethernet and FC. This use case reaps most of the value proposition of FC over Ethernet, without converging the modular switches used to connect the CBB to the data center's existing Ethernet and FC infrastructures. Figure 5 below depicts a Blade server based CBB fabric convergence.

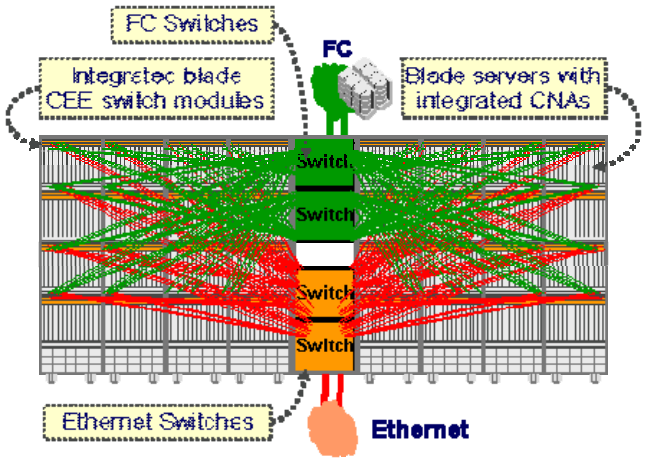

Figure 5

Each server in figure 5 uses a dual-ported Converged Network Adapter (CNA) to connect through CEE to an integrated Blade Server switch, which connects to the CBB's modular Ethernet switches through Ethernet and the modular FC switches through FC. Similar to use cases (A) and (B) above, the CBB doesn't require a change to the data center's Ethernet and FC infrastructure. That is, the Ethernet modular switches within the CBB connect to the data center's existing Ethernet infrastructure. Similarly, the FC modular switches within the CBB connect to the data center's existing FC infrastructure.

Though the CBB shown in figure 5 only depicts a Blade server based deployment, a similar configuration can be constructed with either rack optimized or large SMP servers. For rack optimized servers, FCoE TOR switches connect to the CBB's Ethernet modular switches through Ethernet and to the CBB's FC modular switches through FC.

The CBB level fabric convergence use case has the potential of saving a considerable amount of hardware, through the elimination of dedicated FC equipment within each chassis in the CBB. For example, an eight rack configuration that has 4 blade chassis per rack and 14 server blades per chassis has the potential of eliminating 448 FC adapters and 64 FC integrated blade switches per CBB.

For the same reasons as described in use case (A), we recommend using a development/test environment before deploying the CBB level use case in production.

D. Converged Cloud Building Block

This use case fully converges the fabric within the CBB, but at the data center level there are still two separate fabrics: Ethernet and FC. This use case provides the full value proposition of FC over Ethernet convergence within the CBB, without having to rip and replace the data center's existing Ethernet and FC infrastructures.

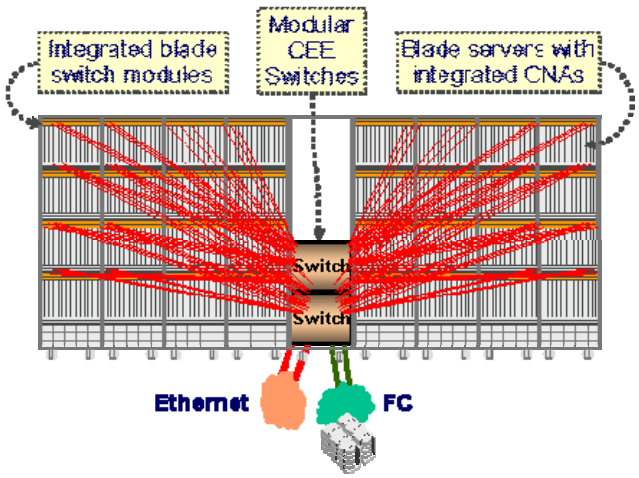

Figure 6

Figure 6 depicts a Blade server based CBB fabric

convergence.

Each server in figure 6 uses a dual-ported Converged Network Adapter (CNA) to connect to the CBB's modular switch through an integrated CEE Blade Server switch.

Similar to use cases (C) above, the CBB connects to the existing data center's 10 Gbps SFP+ Ethernet and FC infrastructure. So, no changes are required to existing data center cabling. The CBB shown in figure 6 depicts a Blade server based deployment, a similar configuration can be constructed using CEE enabled TOR switches with either rack optimized or large SMP servers.

Similar to (C) above, this use case eliminates the Fibre Channel adapters, cables and access switches (i.e. blade switches in figure 6). Within the CBB, it also eliminates the modular FC switches. That is, it uses 2 converged modular switches vs the 2 Ethernet and 2 FC switches shown in use case (C) above.

For the same reasons as described in use case (C), we recommend that use case (D) be deployed in a development/test environment before deploying this CBB level use case.

CBB level fabric convergence can also be used for large SMPs. In this case for the reasons mentioned in use case (B) above, each large SMP's CNAs would likely connect directly to the Modular switch using CEE. The same value proposition and savings described above would apply in this case.

E. Storage attachment to converged fabrics

Each of the uses cases covered so far can take an additional evolutionary step towards full fabric convergence by using FCoCEE based storage servers.

Figure 7 depicts a converged CBB example for the storage attachment use case. In this example, the storage servers within the CBB use FC over CEE capable CNAs to connect into a CEE portion of the fabric.

Using FCoCEE capable CNAs at the servers has the potential for eliminating an adapter at the server. However, migrating from FC attached storage to FCoCEE attached storage doesn't yield the same result, because it just replaces an FC adapter with an FCoCEE capable CNA. Therefore, for most data center environments, migrating FC attached storage to FCoCEE attached storage doesn't yield a capital equipment savings.

As the management stacks for storage and IP/Ethernet mature and are more tightly integrated, thereby allowing a tools consolidation, FCoCEE attached storage has the potential of yielding operational expense savings.

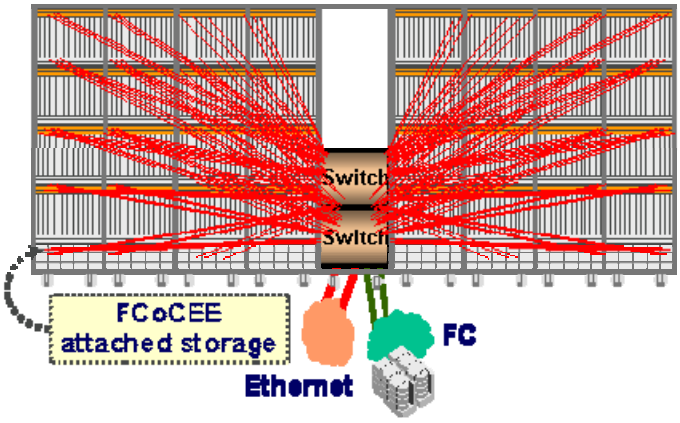

Figure 7

F. FICON level convergence

FICON environments have requirements above most open FC configurations[7]. For example, FICON requires FC Class 2 for faster error detection. Intermediate CEE only switches residing between a server and an FC Forwarder (FCF) would not respond in the event of errors to Class 2 frames as legacy FC switches would. Class 2 frames, just like all other FCoCEE frames, would be treated just as any other Ethernet frame. Any frames that could not be delivered due to congestion or offline destinations would eventually be dropped by the CEE switch, not busied (F_BSY) or rejected (F_RJT) as they would by an FC aware switch. FC-BB-5 makes note of this potential lack of Class 2 functionality, but leaves the solution to managing the supported configurations and simply not allowing intermediate CEE only switches.

Another issue FICON would have with intermediate CEE only switches is Link Incident Detection and Reporting. The FC Back Bone 5 specification (FC-BB-5) is defining a Link Error Status Block definition for Ethernet port statistics to satisfy the reporting requirement. FC-BB-5 has also addressed detection of lost links by periodic Link Keep Alive and Advertisement messages. However, the timeliness of the detection (the period between Keep Alives can be large) coupled with the above mentioned lack of Class 2 responses this could be problematic for FICON, further pointing to the need for no intermediate CEE only switches.

There is also a desire in FICON environments to support direct server to storage attachment configurations. FCoE as defined in FC-BB-5 does not support such a configuration as there must always be an FCF. The current project proposal for FC-BB-6 contains this capability as an item to be

addressed.

G. Data Center level convergence

The ultimate evolutionary step is a fully converged Data Center, where all switches carry FCoCEE traffic. Though this use case can be achieved directly, a more prudent approach is to get to data center level convergence by progressing through either use case (C) or (D). This can be achieved by deploying FCoCEE capable switches throughout the data center, but only using them as Ethernet switches initially. As the organizational silo and fabric management issues described early in this paper are addressed, the data center's FCoCEE switches can either use FC line cards to attach FC storage or simply attach FCoCEE capable storage directly.

III. Conclusion

With the large install base of FC based storage in the enterprise datacenter, FCoCEE offers a fabric convergence solution that aims to protect FC storage investment while providing a consolidated network for clustering, storage and IP/Ethernet traffic. As FCoCEE matures and meets the performance, reliability and quality requirements of Enterprise customers, we expect CEE will play well in large enterprises wanting to pursue FC convergence with Ethernet.

The use cases described in this paper provide an evolutionary model for deployment of FCoCEE based converged fabrics. Just focusing on the potential hardware savings would lead to moving directly to use case G.

As shown in table 1, consideration must also be given to the fabric contention scope associated with each use case and how much of the existing infrastructure is protected. As the fabric contention scope widens, additional Ethernet enhancements beyond the initial set of Convergence Enhanced Ethernet proposal. In our view, two additional enhancements are required[1]: link level congestion notification and a layer-2 multi-pathing mechanism, such as link level shortest path first based routing protocol.

As the scope widens the management infrastructure also needs to be robust enough to provide data center wide planning, configuration and monitoring for each Ethernet traffic class. Fibre Channel convergence with Ethernet also requires the integration of the FC based infrastructure that manages virtual FC (i.e. FCoCEE attached) and physical FC (i.e. natively attached) devices, with the underlying Ethernet infrastructure. In other words, consideration must be given to the maturity of data center wide converged fabric discovery, configuration, monitoring and accounting tools.

In our view, for existing data centers, use cases (A) and (B) provide a natural transition into converged fabrics. These use cases mitigate the organizational and management issues, while enabling significant capital (fewer elements) and operational (Energy efficiency) savings. They also protect the existing infrastructure investment (hardware, management tools, and skills) in Ethernet and Fibre Channel fabrics.

Use case (C) and (D) provide a Cloud Building Block based evolutionary step towards a fully converged data center. Both have the potential for providing large capital and operational savings, while enabling clients to leverage the proven Ethernet and Fibre Channel infrastructure their organizations are familiar and skilled in. Use case (E) provides attachment of storage directly to a CEE fabric, the amount of FC equipment that is eliminated in this case depends on how use case (E) is mixed with use cases (A) through (D). As the FCoCEE technologies, products and the management infrastructure mature, we expect the additional use cases to be pursued.

		A.1	A.2	B	C	D	E	F	G
What native Fibre Channel equipment is eliminated within each use case installation?	Server Adapters	✓	✓	✓	✓	✓	✓		✓
	Server Cables	✓	✓	✓	✓	✓	✓		✓
	Access Switches	✓	✓	✓		✓	✓		✓
	Cloud Cell Modular Switches				✓	✓	✓		✓
	Storage Adapters						✓		✓
	Storage cables						✓		✓
	System z Adapters							✓	✓
	System z cables							✓	✓
	Data Center Modular Switches								✓
What is the fabric contention scope for each use case?	Adapter to Access	✓	✓	✓	✓	✓	✓	✓	✓
	Access to Modular					✓			✓
	Full DC								✓
Is existing DC infrastructure protected?		✓	✓	✓	✓	✓	✓	✓	

Table 1

Acknowledgements

The authors are grateful to Scott Carlson, Rob Cowart, Roger Hathorn, Mike Kaczmarski, David Kahn, James Macon, Manoj Wadekar, Jeff Palm, Dhabaleswar Panda and Suresh Vobbilisetty, for their insightful comments.

References

[1] M. Ko, D. Eisenhauer, R. Recio, "A Case for Convergence Enhanced Ethernet: Requirements and Applications", 2008 IEEE International Conference on Communications

[2] IBM FCoCEE CNA "Brocade 10Gb CNA for IBM System x" http://www.redbooks.ibm.com/abstracts/tips0718.html and "QLogic 10Gb CNA for IBM System x" http://www.redbooks.ibm.com/abstracts/tips0720.html

[3] Brocade, “Brocade 8000 Switch Data Sheet” http://www.brocade.com/downloads/documents/data_sheets/product_data_sheets/8000_DS_00.pdf

[4] W. J. Armstrong, R. L. Arndt, D. C. Boutcher, R. G. Kovacs, D. Larson, K. A. Lucke, N. Nayar, R. C. Swanberg, “Advanced virtualization capabilities of POWER5 systems”, IBM J. RES. & DEV. Vol. 49 No. 4/5 July/September 2005

[5] L. Agarini, G Anselmi, “Integrated Virtual Ethernet Adapter, Technical Overview and Introduction”, IBM Redpaper, October 2007

[6] R. Recio, O. Cardona, “Automated Ethernet Virtual Bridging”, DC CAVES 2009 Workshop, collocated with the 21st International Tele-traffic Congress (ITC 21)

[7] S. Carlson, SB-3 Concerns with FCoE, presented at FC-BB-5 10/08 meeting, see http://www.t11.org/ftp/t11/pub/fc/bb-5/08-590v0.pdf